Jagua Publication

Computational Methods, 3(2):27–34, July-December: 2026.
DOI: https://doi.org/10.58614/cm324

Computational Methods
ISSN (Online): 3048-7447

Journal homepage: https://www.jpub.org/journal-details.php?journal-id=43

Research Article

A Predictive Risk Framework for P2P Micro-Lending Default Prediction with Anomaly Detection

Christopher Ekong [1], Aniefiok Ukommi [2], Blossom Okorie[2], Bliss Utibe-Abasi Stephen[3*] and Philip Asuquo[3]

[1]*Department of Economics, University of Uyo, Uyo, Nigeria.*
[2]*University of Uyo, Uyo, Nigeria.*
[3]*TETFund Centre of Excellence in Computational Intelligence Research, University of Uyo, Uyo, Nigeria.*
[*]*Corresponding author: blissstephen@uniuyo.edu.ng*

Article Info

Keywords: *Peer-to-peer lending, Default prediction, LightGBM, SHAP explainability, Anomaly detection.*

Received: *05.08.2026;*
Accepted: *02.09.2026;*
Published: *07.09.2026*

Abstract

Machine-learning credit assessment is often discussed as a substitute for manual underwriting, yet high-stakes lending decisions require a more careful design in which predictive models augment human judgement. This paper presents a supervised learning framework for peer-to-peer micro-lending default prediction using a reproducible pipeline built around LightGBM classification, engineered affordability features, class-imbalance handling, Isolation Forest anomaly detection, and SHAP explainability. The dataset is processed through exploratory analysis, feature construction, stratified train-test splitting, median and modal imputation, robust scaling, categorical encoding, SMOTE rebalancing, model training, diagnostic evaluation, business-impact estimation, and global and local explanation. The experimental run reports an AUC-ROC of 0.7459, average precision of 0.3042, macro-F1 of 0.5614, and default-class F1 of 0.3332. At the operating threshold used in the pipeline, the model identifies 4,081 defaults while missing 1,850 defaults, representing $236,020,901.20 in potential principal at risk. SHAP analysis ranks age, interest rate, months employed, credit-to-income, dependents, and co-signer status among the strongest drivers, revealing both useful affordability signals and governance-sensitive proxy variables. The findings support an augmentation thesis: AI can improve default-risk triage, anomaly surfacing, and explanation consistency, but false positives, missed defaults, proxy-feature risks, and high-value edge cases require human-in-the-loop underwriting, fairness review, and accountable lending policy.

1. Introduction

Credit allocation is a consequential decision process because it determines which borrowers can access liquidity, which lenders bear default losses, and which communities are included or excluded from financial opportunity. Machine-learning credit-scoring research shows that supervised models can process high-volume applications, capture nonlinear borrower patterns, and improve default-risk discrimination compared with traditional linear scorecards [1]. At the same time, algorithmic lending research warns that better prediction is not identical to better governance: automated credit systems may reproduce historical bias, intensify information asymmetry, or create adverse outcomes that are difficult for applicants to understand or contest [2, 3].

This paper addresses a practical question: when a supervised credit-risk model is available, should it replace human underwriting or augment human decision-making? The answer developed here is augmentation. A model can rank applications, surface latent risk patterns,

identify anomalous records, and generate explanation artifacts for reviewers. However, the same model can also produce false positives, miss high-value defaults, and rely on features whose social meaning requires policy judgement. In credit decisioning, these residual risks are not peripheral. They define the boundary between responsible automation and unsupported delegation.

The primary experimental artifact is a complete micro-lending default-risk classification pipeline whose modules cover data loading, exploratory analysis, feature engineering, preprocessing, imbalance treatment, LightGBM model training, evaluation, and SHAP explanation [4]. The pipeline uses boosted trees for heterogeneous tabular credit data, handles default-class imbalance through SMOTE by default, provides alternative imbalance modes through undersampling and class weighting, detects outlying training records with an Isolation Forest, and saves both aggregate and local explanation outputs. This makes it suitable for studying not only predictive performance but also human-in-the-loop governance.

The resulting performance is useful but imperfect. The evaluation reports AUC-ROC of 0.7459 and default recall of 0.69, indicating that the classifier captures many risky loans. Yet default precision is only 0.22, meaning that many applicants predicted as defaulting are actually safe. The model also misses 1,850 defaults, producing a principal-at-risk estimate of $236.02 million. These results make replacement implausible and augmentation necessary. They also demonstrate why aggregate accuracy alone is not a sufficient measure for lending systems: error direction, economic exposure, and borrower fairness must be analysed together.

The central contribution of this paper is an augmentation framework for credit-risk AI. The framework has three components. First, predictive augmentation uses the LightGBM model to triage applications by estimated default risk. Second, explanatory augmentation uses SHAP outputs to identify the features most responsible for both global model behaviour and individual predictions. Third, governance augmentation converts anomalies, borderline probabilities, proxy-sensitive features, and high-exposure errors into mandatory human-review triggers. The paper therefore documents a reproducible supervised-learning workflow, converts its empirical outputs into operational evidence, and proposes a review architecture in which model scores, explanations, anomaly flags, and underwriter decisions are logged as a single accountable process.

2. Related Work

2.1. Machine Learning for Loan Default Prediction

Credit-risk literature has moved from purely statistical scorecards toward boosted, hybrid, and deep-learning models. Hybrid LightGBM studies report that convolutional feature extraction combined with gradient-boosted trees can outperform several classical comparators in loan-default prediction [5]. Explainable default-prediction research comparing logistic regression, decision trees, XGBoost, and LightGBM also finds that boosted-tree models improve predictive performance while requiring post-hoc explanation for usability and trust [6]. Other LightGBM-based credit models report strong results against XGBoost, Lasso, and CatBoost comparators in structured default-prediction settings [7]. These findings support the use of LightGBM as the main classifier in the present framework.

The strongest credit-risk systems integrate model choice with domain-aware feature engineering, imbalance treatment, and explainability. Enhanced LightGBM frameworks for digital finance motivate their designs around high-dimensional data, minority-class skew, and limited interpretability [8]. Analytical credit-risk systems also use SHAP to convert model output into actionable risk evidence for banking decision-makers [9]. The present framework is simpler than advanced hybrid architectures, but it follows the same operational logic: structured feature construction, boosted classification, class-imbalance handling, and explanation artifacts that can be inspected by humans.

2.2. Class Imbalance, Explainability, and Governance

Default prediction is an imbalanced classification problem because most borrowers do not default. This imbalance creates two linked difficulties: aggregate metrics can appear acceptable while default-class performance remains weak, and interpretation methods can become unstable when the minority class is rare. Studies of imbalanced credit-scoring datasets show that class skew can affect the stability of LIME and SHAP explanations [10]. Oversampling research for explainable credit scoring proposes non-parametric methods to improve minority-class learning while retaining interpretability [11, 12].

Explainability is central in credit scoring because adverse decisions must be contestable, auditable, and understandable to operational staff. Research on automated credit decisions argues that black-box models undermine borrower and lender trust and uses SHAP to identify feature importance and examine the effect of removing potentially discriminatory variables [13]. Fairness studies in banking credit scoring show that historical data can encode discrimination and that bias-mitigation methods must be evaluated alongside predictive performance [14]. Broader work on algorithmic discrimination in credit shows how models trained on past decisions can consolidate bias against protected or socially salient groups [3]. Human-AI collaboration research also shows that decision control, uncertainty communication, and stakeholder-specific explanation design influence whether users can meaningfully govern AI systems [15–17].

3. Methodology

3.1. Dataset and Exploratory Analysis

The study uses a loan-default dataset with a binary target where 1 denotes default and 0 denotes safe repayment. The data include borrower and loan attributes such as loan amount, loan term, income, months employed, age, debt-to-income ratio, interest rate, employment type, loan purpose, credit score, number of credit lines, dependents, co-signer status, and mortgage status [4]. Exploratory analysis checks shape, data types, missingness, class balance, and correlations, then saves diagnostic plots for interest rate, debt-to-income ratio, credit-score distribution, and numeric feature relationships.

Figure 1 links the main distribution diagnostics to the modelling problem. Interest rate and debt-to-income ratio are direct affordability and risk indicators, while the credit-score histogram compares the safe and default classes across a conventional creditworthiness measure. Figure 2 complements these plots by showing linear relationships among numeric variables, which helps identify redundant variables and motivates nonlinear tree-based modelling when simple correlations are insufficient.

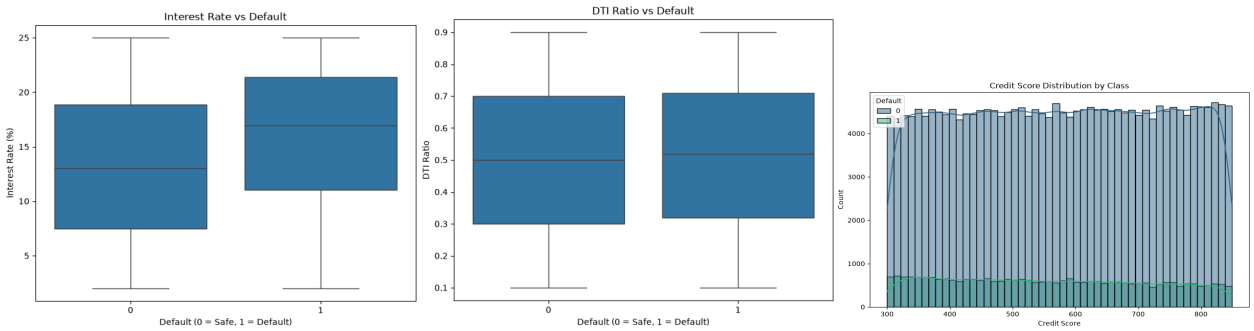

Figure 1: Exploratory distribution diagnostics: interest rate by default class, debt-to-income ratio by default class, and credit-score distribution by default class.

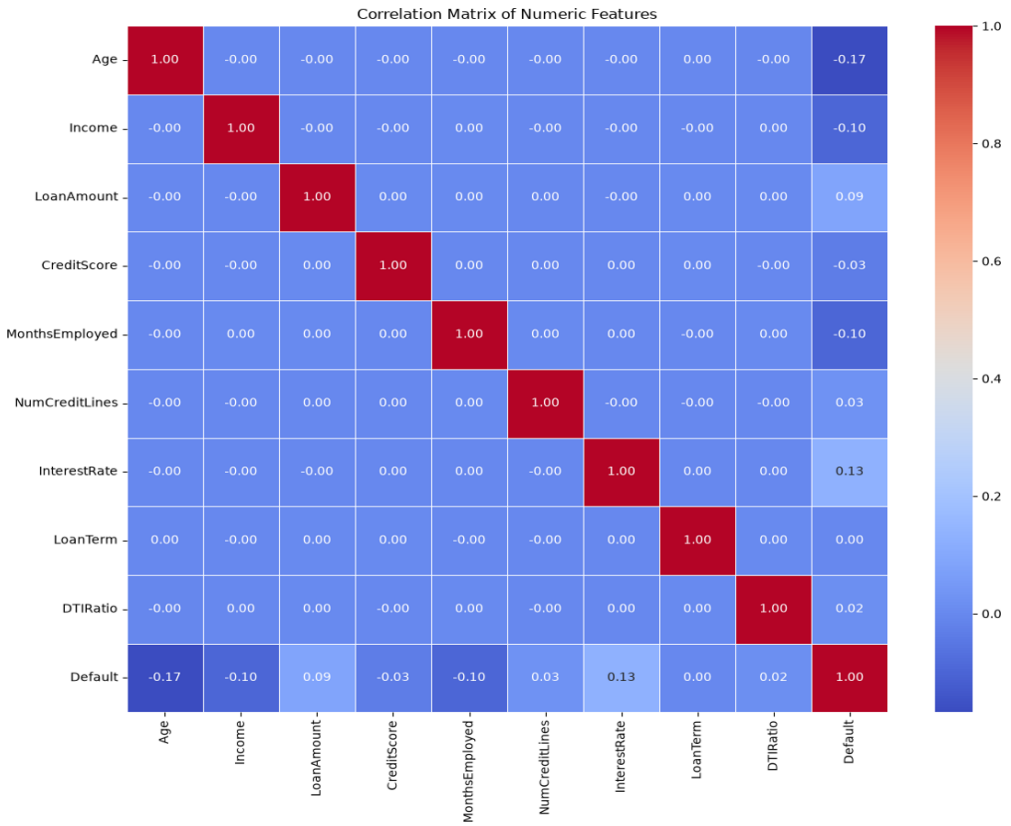

Figure 2: Correlation matrix for numeric borrower and loan variables.

3.2. Feature Engineering and Preprocessing

Five domain features are engineered before modelling. Payment-to-income divides estimated monthly loan payment by monthly income. Credit-to-income divides the loan amount by annual income. Employment-age ratio relates months employed to applicant age in months. High-DTI flag marks debt-to-income ratios above 35. Interest-rate tier bins interest rates into ordinal risk bands. These features translate raw application fields into underwriting concepts such as affordability, leverage, employment stability, and risk-based pricing.

The pipeline uses a stratified 80/20 train-test split before imputation to reduce leakage. Numeric variables are imputed with training medians, while categorical variables are imputed with training modes. Employment type and interest-rate tier are label encoded, and remaining categorical variables are one-hot encoded with first-level dropping and unknown-category handling. Income and loan amount are robust-scaled and log-transformed after applying a nonnegative offset. This design preserves the test set as an unseen evaluation sample and avoids fitting imputers, scalers, or encoders on the full dataset.

3.3. Class Imbalance, Model Training, and Anomaly Detection

The minority default class is handled through three selectable strategies. The default strategy applies SMOTE with five neighbours to synthetically increase default-class examples. A second strategy applies random undersampling to reduce the majority safe class. A third strategy leaves the training data unchanged but passes a class-weight ratio to the LightGBM classifier. The pipeline also applies an Isolation Forest with a five-percent contamination setting to identify anomalous training rows. These outliers can be saved for review or optionally excluded before training [4].

The classifier is a LightGBM model configured with 1,000 estimators, a learning rate of 0.05, 63 leaves, unrestricted depth, minimum child samples of 20, row subsampling of 0.8, column subsampling of 0.8, L1 and L2 regularisation of 0.1, and a fixed random seed. Optional

Bayesian hyperparameter search uses five-fold stratified cross-validation and AUC-ROC scoring over number of leaves, learning rate, maximum depth, and minimum child samples.
The business-impact calculation is defined as

$$R_{FN} = N_{FN}\bar{L} = 1850 \times 127578.87 = 236020901.20, \tag{1}$$

where R_{FN} is potential principal at risk from false negatives, N_{FN} is the number of missed defaults, and $\bar{L}$ is the implied average exposure per missed default.

4. Results and Analysis

4.1. Model Performance

Table 1 summarises the core evaluation results. AUC-ROC of 0.7459 indicates meaningful separation between safe and defaulting borrowers, while average precision of 0.3042 reveals the difficulty of precision under minority-class imbalance. Macro-F1 of 0.5614 is substantially lower than accuracy, showing that the safe and default classes are not equally easy to classify. Default-class F1 of 0.3332 is the strongest evidence against autonomous replacement: the model is useful for risk triage, but not reliable enough to make final adverse decisions without review.

Table 1: Evaluation metrics for the default prediction model.

Metric	Value
AUC-ROC	0.7459
Average precision	0.3042
Macro-F1	0.5614
Default-class F1	0.3332
Accuracy	0.68
Test observations	51,070

Figure 3 links the numerical performance measures to their diagnostic plots. The ROC curve summarises threshold-independent separation, the precision-recall curve highlights the minority-class default challenge, and the confusion matrix shows how true and false decisions are distributed at the operating threshold.

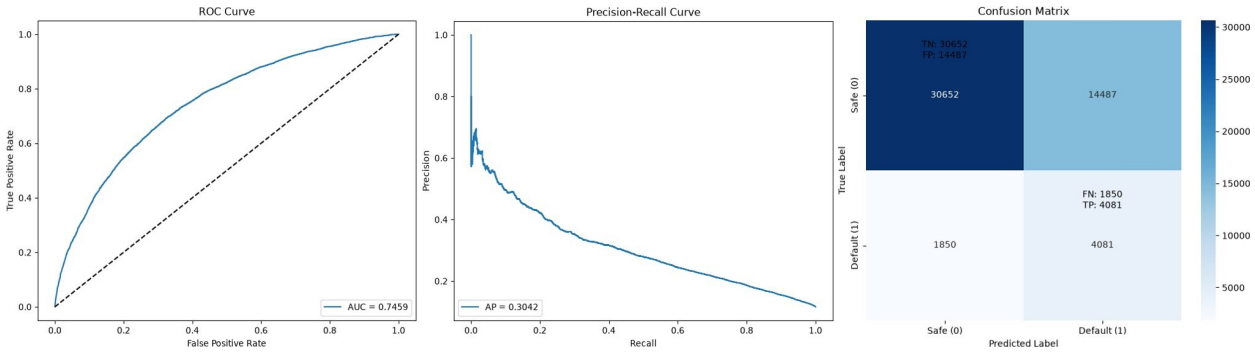

Figure 3: Evaluation diagnostics: ROC curve, precision-recall curve, and confusion matrix for safe and default classes.

4.2. Classification Errors and Business Impact

Table 2 provides the class-level results. The model achieves high precision for safe borrowers, 0.94, but only 0.22 precision for defaults. Conversely, default recall is 0.69, which means the model captures most observed defaults but at the cost of 14,487 false-positive default predictions. A borrower falsely classified as likely to default may face denial, higher scrutiny, or less favourable terms. Thus, false positives are not harmless errors; they are cases where underwriters should inspect explanations, affordability evidence, and policy constraints before acting.

Table 2: Classification report and confusion matrix counts.

Class	Precision	Recall	F1	Support
Safe	0.94	0.68	0.79	45,139
Default	0.22	0.69	0.33	5,931
Confusion matrix counts				
TN		30,652		
FP		14,487		
FN		1,850		
TP		4,081		

The business-impact estimate in Table 3 reframes the same confusion matrix in financial terms. False negatives represent defaulting borrowers predicted as safe. The run reports 1,850 missed defaults, equal to $236,020,901.20 in potential principal at risk. Dividing the principal-at-risk estimate by the number of false negatives implies an average exposure of approximately $127,578.87 per missed default.

Table 3: Business impact of missed defaults.

Business measure	Value
False negatives	1,850
Potential principal at risk	$236,020,901.20
Implied exposure per missed default	$127,578.87
True defaults detected	4,081
False-positive default predictions	14,487

4.3. Explainability Results

Figure 4 presents the global SHAP diagnostics. The bar chart ranks mean absolute feature contribution, while the beeswarm plot shows how feature values push individual predictions toward or away from default risk. Together, these plots reveal the model's operational logic: age, interest rate, employment duration, credit-to-income, dependents, co-signer status, employment type, credit score, number of credit lines, and log loan amount dominate global behaviour.

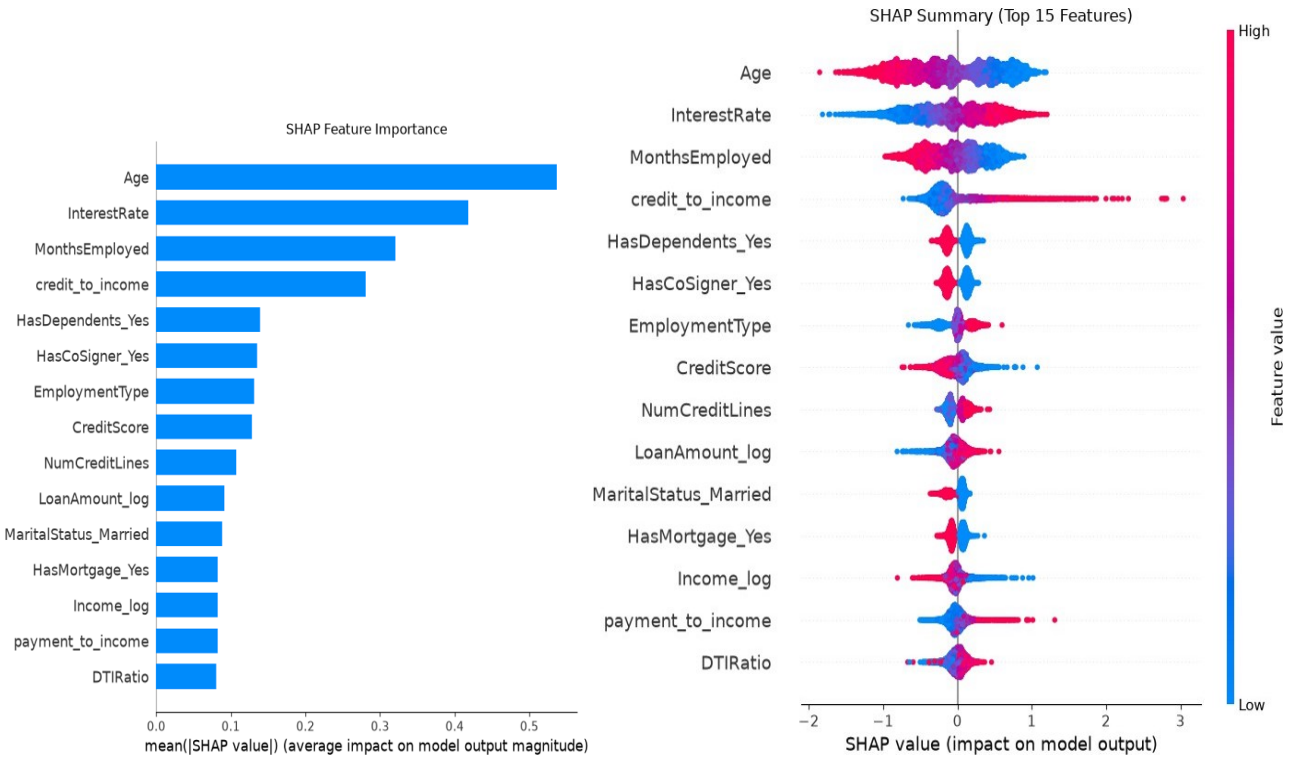

Figure 4: Global SHAP diagnostics: mean absolute feature importance and beeswarm distribution of feature contributions across sampled test records.

Table 4 converts the global SHAP ranking into governance interpretation. Some features are directly interpretable affordability signals: interest rate, credit-to-income, loan amount, and DTI ratio describe repayment burden. Others require governance caution. Age, dependents, marital status, and co-signer status may reflect repayment patterns, but they may also act as proxies for life stage, family structure, or protected characteristics depending on jurisdiction and policy context.

Table 4: Top SHAP-ranked features and governance interpretation.

Rank	Feature	Governance interpretation
1	Age	Strong signal; fairness-sensitive and requires policy justification.
2	InterestRate	Direct loan-risk and pricing signal.
3	MonthsEmployed	Employment stability and income-security proxy.
4	credit_to_income	Affordability and leverage signal.
5	HasDependents_Yes	Household-obligation proxy; review for fairness.
6	HasCoSigner_Yes	Risk-sharing and guarantee signal.
7	EmploymentType	Labour-market stability signal.
8	CreditScore	Conventional creditworthiness signal.
9	NumCreditLines	Credit exposure and history signal.
10	LoanAmount_log	Exposure-size signal after robust scaling.

Figure 5 shows local SHAP waterfall diagnostics for three underwriter-relevant cases: the highest-confidence default, the highest-confidence safe prediction, and the most borderline prediction. These plots connect each individual decision to its strongest feature-level reasons.

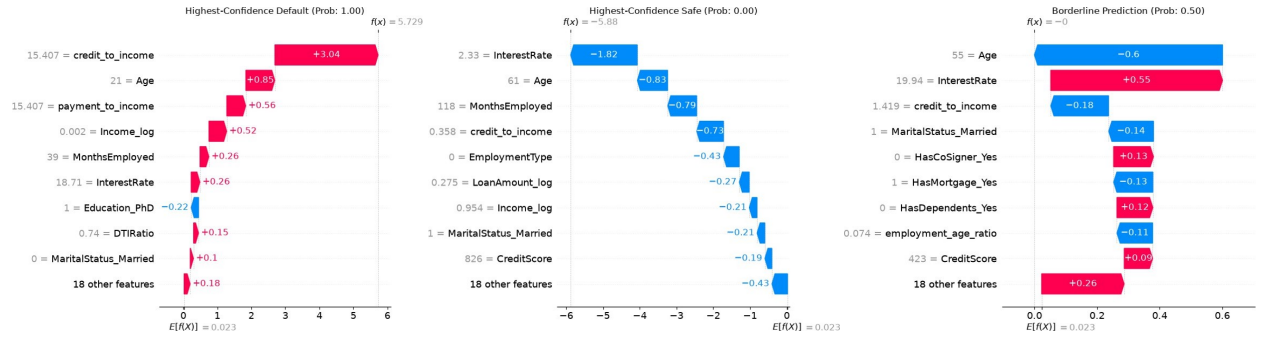

Figure 5: Local SHAP waterfall diagnostics: highest-confidence default case, highest-confidence safe case, and borderline case closest to the decision threshold.

5. Governance Discussion

The results strongly support the augmentation thesis. If the evidence had shown near-perfect classification, one might argue for a more automated decision pathway. Instead, the model reports moderate discrimination, strong default recall relative to the baseline prevalence, weak default precision, and substantial false-negative exposure. This combination is typical of useful decision support in high-stakes domains. The AI system improves the underwriter's ability to find risk, but it also creates new review obligations.

The first implication is that evaluation must be business-aware. Accuracy of 0.68 is not the central result. The central result is the distribution of errors. A false negative exposes the lender to principal loss, while a false positive exposes the borrower to possible denial or unfavourable treatment. The report quantifies only the lender-side cost of false negatives. A production system should also estimate borrower-side costs, such as rejected safe borrowers, delayed decisions, or unnecessary manual requests for documentation. This would turn Table 3 into a two-sided risk ledger.

The second implication is that explainability is not merely a compliance artifact. The SHAP ranking reveals the model's operational logic. Interest rate, employment duration, credit-to-income, credit score, and number of credit lines are intuitive credit-risk factors. Age, dependents, marital status, and co-signer status are more delicate. They may be statistically predictive, but they require policy review because they can encode demographic or household-structure effects. In this framework, SHAP functions as a governance sensor: it does not automatically certify fairness, but it tells reviewers where to look.

The third implication is that anomaly detection defines an irreducible human role. The Isolation Forest layer identifies approximately five percent of training observations as anomalous by configuration. These cases are precisely where model generalisation is least secure. They may include data-entry errors, unusual borrower profiles, fraud patterns, or legitimate edge cases underrepresented in the training distribution. Automated rejection of such cases would confuse anomaly with default risk. A better workflow routes anomaly cases to humans with model score, SHAP attribution, raw application data, and reason codes visible in one review interface.

Several limitations remain. The archive does not include a temporal validation design, so the reported performance may not reflect drift across credit cycles. The business-impact metric uses mean loan amount rather than realised loss given default, collateral recovery, or exposure at default. SMOTE improves minority-class learning but may affect calibration, and the archive does not report calibration curves or threshold optimisation. The fairness analysis is based on SHAP visibility rather than formal group fairness metrics, counterfactual tests, or adverse-impact analysis. Finally, the dataset is public and generic, so deployment in a real lender would require data-governance review, privacy assessment, adverse-action explanation design, and external validation. As shown is Table 5.

Table 5: Human-in-the-loop governance triggers.

Trigger	Required human action
Borderline probability near 0.5	Underwriter review before approval or rejection.
Isolation Forest anomaly	Check data quality, fraud indicators, and unusual borrower context.
High exposure amount	Senior credit review because false negatives are financially material.
SHAP dominated by age or family proxies	Fairness and policy justification review.
False-positive-sensitive segment	Inspect affordability evidence before adverse decision.
Model-rule disagreement	Escalate to policy owner and document override rationale.

6. Conclusions

This paper presents a supervised LightGBM framework for peer-to-peer micro-lending default prediction with engineered affordability features, SMOTE imbalance handling, Isolation Forest anomaly detection, diagnostic evaluation, business-impact analysis, and SHAP explainability; the empirical results show that the model provides useful risk triage by detecting 4,081 defaults, but it also misses 1,850 defaults representing $236.02 million in potential principal at risk and produces many false-positive default predictions, so the appropriate deployment model is not autonomous replacement of underwriters but accountable augmentation in which scores, explanations, anomaly flags, high-exposure cases, and proxy-sensitive features are routed through human review, fairness oversight, and documented lending policy.

Article Information

Funding: No external funding was received for this study.

Author Contributions: All authors contributed to the conception, methodology, analysis, and preparation of the manuscript and approved the final version.

Conflict of Interest: The authors declare no conflict of interest.

Data Availability Statement: Data available on reasonable request.

Disclaimer (Artificial Intelligence): The author(s) hereby declare that NO generative AI technologies such as Large Language Models (ChatGPT, COPILOT, etc.), and text-to-image generators have been used during writing or editing of manuscripts.

References

[1] Helmi Ayari, Ramzi Guetari, and Naoufel Kraïem. Machine learning powered financial credit scoring: A systematic literature review. *Artificial Intelligence Review*, 59(13), 2026. doi: 10.1007/s10462-025-11416-2. URL https://doi.org/10.1007/s10462-025-11416-2.

[2] Holli Sargeant. Algorithmic decision-making in financial services: Economic and normative outcomes in consumer credit. *AI and Ethics*, 3:1295–1311, 2023. doi: 10.1007/s43681-022-00236-7. URL https://doi.org/10.1007/s43681-022-00236-7.

[3] Ana Cristina Bicharra Garcia, Marcio Gomes Pinto Garcia, and Roberto Rigobon. Algorithmic discrimination in the credit domain: What do we know about it? *AI & Society*, 39:2059–2098, 2024. doi: 10.1007/s00146-023-01676-3. URL https://doi.org/10.1007/s00146-023-01676-3.

[4] Project Pipeline Artifact. P2p micro-lending default risk classification pipeline. Primary research artifact supplied with this paper, 2026.

[5] Qiliang Zhu, Wenhao Ding, Mingsen Xiang, Mengzhen Hu, and Ning Zhang. Loan default prediction based on convolutional neural network and LightGBM. *International Journal of Data Warehousing and Mining*, 19(1), 2023. doi: 10.4018/IJDWM.315823. URL https://doi.org/10.4018/IJDWM.315823.

[6] Xu Zhu, Qingyong Chu, Xinchang Song, Ping Hu, and Lu Peng. Explainable prediction of loan default based on machine learning models. *Data Science and Management*, 2023. doi: 10.1016/j.dsm.2023.04.003. URL https://doi.org/10.1016/j.dsm.2023.04.003.

[7] Zhiren Gan, Junyuan Qiu, Fuli Li, and Qian Liang. A LightGBM based default prediction method for american express. In *Proceedings of the 2nd International Conference on Information Economy, Data Modeling and Cloud Computing*, 2023. doi: 10.4108/eai.2-6-2023.2334590. URL https://doi.org/10.4108/eai.2-6-2023.2334590.

[8] Chengwei Ying, Anlu Shi, and Xiongyi Li. Hybrid boosted attention-based LightGBM framework for enhanced credit risk assessment in digital finance. *Humanities and Social Sciences Communications*, 12(1036), 2025. doi: 10.1057/s41599-025-05230-y. URL https://doi.org/10.1057/s41599-025-05230-y.

[9] Marcos R. Machado, Daniel Tianfu Chen, and Joerg R. Osterrieder. An analytical approach to credit risk assessment using machine learning models. *Decision Analytics Journal*, 16:100605, 2025. doi: 10.1016/j.dajour.2025.100605. URL https://doi.org/10.1016/j.dajour.2025.100605.

[10] Yujia Chen, Raffaella Calabrese, and Belen Martin-Barragan. Interpretable machine learning for imbalanced credit scoring datasets. *European Journal of Operational Research*, 312(1):357–372, 2024. doi: 10.1016/j.ejor.2023.06.036. URL https://doi.org/10.1016/j.ejor.2023.06.036.

[11] Seongil Han, Haemin Jung, Paul D. Yoo, Alessandro Provetti, and Andrea Cali. NOTE: Non-parametric oversampling technique for explainable credit scoring. *Scientific Reports*, 14(26070), 2024. doi: 10.1038/s41598-024-78055-5. URL https://doi.org/10.1038/s41598-024-78055-5.

[12] Seongil Han and Haemin Jung. NATE: Non-parametric approach for explainable credit scoring on imbalanced class. *PLOS ONE*, 19 (12):e0316454, 2024. doi: 10.1371/journal.pone.0316454. URL https://doi.org/10.1371/journal.pone.0316454.

[13] Chioma Ngozi Nwafor, Obumneme Nwafor, and Sanjukta Brahma. Enhancing transparency and fairness in automated credit decisions: An explainable novel hybrid machine learning approach. *Scientific Reports*, 14(25174), 2024. doi: 10.1038/s41598-024-75026-8. URL https://doi.org/10.1038/s41598-024-75026-8.

[14] Charlene Mariscal, Yoga Yustiawan, Fauzy Caesar Rochim, and Evawaty Tanuar. Implementing and analyzing fairness in banking credit scoring. *Procedia Computer Science*, 2024. doi: 10.1016/j.procs.2024.03.150. URL https://doi.org/10.1016/j.procs.2024.03.150.

[15] Monika Westphal, Michael Vössing, Gerhard Satzger, Galit B. Yom-Tov, and Anat Rafaeli. Decision control and explanations in human-AI collaboration: Improving user perceptions and compliance. *Computers in Human Behavior*, 144:107714, 2023. doi: 10.1016/j.chb.2023.107714. URL https://doi.org/10.1016/j.chb.2023.107714.

[16] Daniel N. Cassenti, Lance M. Kaplan, and Aayushi Roy. Representing uncertainty information from AI for human understanding. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, volume 67, pages 177–182, 2023. doi: 10.1177/21695067231193649. URL https://doi.org/10.1177/21695067231193649.

[17] Robert R. Hoffman, Shane T. Mueller, Gary Klein, Mohammadreza Jalaeian, and Connor Tate. Explainable AI: Roles and stakeholders, desirements and challenges. *Frontiers in Computer Science*, 5, 2023. doi: 10.3389/fcomp.2023.1117848. URL https://doi.org/10.3389/fcomp.2023.1117848.